Principles of Parametric Estimation in Modeling Language Competition

Menghan Zhang * **and Tao Gong** †

*Research Center for Computational Linguistics, Comparative Linguistics E-Institute, Shanghai Normal University, Shanghai, China, and †Department of Linguistics, University of Hong Kong, Hong Kong

Submitted to Proceedings of the National Academy of Sciences of the United States of America

It is generally difficult to define reasonable parameters and interpret their values in mathematical models of social phenomena. Rather than directly fitting abstract parameters against empirical data, we should define some concrete parameters to denote the socio-cultural factors relevant for particular phenomena, and compute the values of these parameters based upon the corresponding empirical data. Taking the example of modeling studies of language competition, we propose a language diffusion principle and two language inheritance principles to compute two critical parameters, namely the impacts and inheritance rates of competing languages, in our language competition model derived from the Lotka-Volterra competition model in evolutionary biology. These principles assign explicit sociolinguistic meanings to those parameters and calculate their values from the relevant data of population censuses and language surveys. Using four examples of language competition, we illustrate that our language competition model with thus-estimated parameter values can reliably replicate and predict the dynamics of language competition and it is especially useful in cases lacking direct competition data.

language diffusion principle | language inheritance principles | lexical diffusion dynamics

How to define informative parameters in mathematical models of real world phenomena remains as a tough problem; in particular, how to assign explicit meanings to parameters and interpret their values in models of social phenomena critically affects the explanatory power of these models. This issue becomes more serious in recent modeling studies of language dynamics [1, 2, 3, 4, 5], especially competition (the process whereby local tongues are being replaced by hegemonic languages due to population migration and socio-cultural exchange [6]).

Among the numerous modeling approximations of two-language competition [7, 8, 9, 10, 11, 12, 13, 14, 15], the most influential one was the AS model [8]. It defined prestige (the socio-economic status of the speakers of a language) of competing languages to determine the dynamics of language competition, and reported well-fitting curves to some historical data under a fixed range of prestige value. However, this abstract parameter lacked explicit socio-cultural meanings; it remained unclear what were the characteristics of a language having a prestige value, say 1.2, and what was the socio-cultural condition corresponding to the difference between two languages having prestige values, say 1.2 and 1.3, respectively. Lacking such empirical foundations, the prestige value had to be obtained via curve fitting, thus making this model useless in cases lacking sufficient empirical data. Although many recent models [9, 10, 11, 12, 13, 14, 15] extended the AS model in certain aspects (e.g. the MP model [9] incorporated bilinguals into competition, the SS model [10] adopted network structures to confine language contact, and the MW model [11] revealed the possibility of preserving endangered languages by enhancing their relatively-small prestige values), most of them kept using prestige in their discussions of language competition and pertinent issues. Language competition is subject to many socio-cultural constraints, among which the primary ones include: the population sizes of competing languages, the geographical distances between these populations, and the non-uniform population distributions in competing regions [5, 13, 16, 17, 18, 19, 20]. Prestige alone fails to explicitly address these many factors, and applying fixed prestige values in different cases of language competition apparently disregards the actual conditions of those cases.

Noting these, we define two concrete parameters, namely the impacts and inheritance rates of competing languages, and adopt the Lotka-Volterra competition model [21, 22, 23] in evolutionary biology to study the dynamics of language competition. Meanwhile, we propose a language diffusion principle and two language inheritance principles to calculate these parameters based on the relevant data of population censuses and language surveys. The language diffusion principle, inspired by the Fourier's law of heat conduction, computes the impacts of competing languages from the population sizes of these languages and the geographical distances between the region where competition occurs and the population centers of these languages. The empirical data for this calculation are available in population censuses and geographical information systems. The language inheritance principle I, inspired by the Hardy-Weinberg genetic inheritance principle [24, 25], computes the inheritance rates of competing languages based on the occurring frequencies of these languages during language learning. Both monolinguals and bilinguals are taken into account, and the empirical data for this calculation can be extracted from the surveys of speakers' language choices in communities. The language inheritance principle II, modified from the well-attested lexical diffusion dynamics [26, 27], adopts the logistic curve [29] to estimate the inheritance rates of competing languages. This makes the principle applicable in cases lacking sufficient data of speakers' language choices. Following these principles, the calculated parameter values can clearly indicate the influence of those primary factors on language competition. Based on our language competition model, in practice, rather than curve fitting, we first explicitly compute the values of these parameters, and then, use our model with thus-estimated parameter values to replicate the dynamics of language competition in particular cases of language competition. Based on the language inheritance principle II, our model can also reasonably predict the dynamics of language competition in cases that lack direct competition data.

Materials and Methods

Language competition model. When multiple languages come into contact, one or more of them may become endangered, due to the fact that speakers may prefer using the other of them. Such competition can be viewed as a process that these languages gain survival advantage via resource plunder. Resource here refers to the speakers in the competing region, the survival advantage of

Reserved for Publication Footnotes

a language manifests primarily in its number of speakers in this region, and the competition dynamics is reflected mainly by the change in the population sizes of these languages in this region. On these aspects, language competition resembles the competing relation in ecology, where the rise or decline of the population size of a species is influenced by the growth rate of the competing species. This competing relation exists not only between predators and preys, but common among various species in the biological world. In evolutionary biology, the Lotka-Volterra competition model (the original form was proposed to describe the predator-prey competition [21, 22], but its generalized form [23] could also examine the general competition among various species and trace its dynamics) has been proved to be able to reliably describe the dynamics of such competition. Therefore, we derive our language competition model from this well-attested model, and assign linguistic meanings to its parameters in order to fit it into the situation of language competition.

Our macroscopic model consists of two first-order differential equations, which denote the conversion functions describing the change in populations speaking two competing languages **[1]**:

$$\begin{cases} \frac{dx_1}{dt} = r_1 x_1 (1 - \frac{x_1}{N_1} - \sigma_1 \frac{x_2}{N_2}) \\ \frac{dx_2}{dt} = r_2 x_2 (1 - \frac{x_2}{N_2} - \sigma_2 \frac{x_1}{N_1}) \end{cases} \quad \textbf{[1]}$$

Here, $x_1(t)$ and $x_2(t)$ denote the numbers of speakers of two competing languages L_1 and L_2 in a particular region and at a particular time t. N_1 denotes the maximum size of the monolingual population speaking L_1 in this region, and N_2 the maximum size of the monolingual population speaking L_2. σ_1 denotes the impact of L_2 on L_1, and σ_2 the impact of L_1 on L_2. r_1 and r_2 denote the inheritance rates of the populations speaking L_1 and L_2.

Instead of the dynamics of population growth, this model examines how competing languages plunder speakers based on their impacts and inheritance rates. The dynamics of language competition is collectively determined by these two parameters, and how to assign explicit meanings to them and estimate their values based on corresponding data becomes critical for using these parameters to denote the influence of those socio-cultural factors in particular case of language competition.

Language diffusion principle. In our competition model, the impact of a language (σ) refers to the influence of this language on other language(s) in the competing region after this language diffuses into this region. People are language carriers, language diffuses along with the diffusion of the population from the population center of the speakers of this language to the competing region, and with the increase in the distance between the population center and the competing region, the impact of this language decreases.

We propose a language diffusion principle to calculate the impacts of competing languages. It is inspired by the Fourier's law of heat conduction. We assume that (a) the center of the population speaking a particular language has the maximum population density (this may not often hold in reality, due to historical, political or economic reasons; we need to estimate such 'population center' based on population density in particular cases); (b) the geographical distance is inversely proportional to the population size: the further the distance from the center, the smaller the number of individuals (this is more valid in early times, or in populations not living in developed states with a long history of spatial structuration); (c) the population diffusion occurs in all directions at the same rate, regardless of disturbance from ecological factors or social policy; and more importantly, (d) the population diffusion follows the Fourier's law of heat conduction. Following these assumptions, we define the population diffusion principle as in equation **[2]**:

$$C(d,t) = \frac{Q}{(4\pi kt)^{\frac{3}{2}}} e^{-\frac{d^2}{4kt}} \quad \textbf{[2]}$$

Here, in an unlimited 2D space, at time t and a particular region (x, y) where competition takes place, d is the Euclidean distance from the origin of coordinates (0, 0) to this region, Q is the population size at the center, k is the constant diffusion coefficient, and C calculates the ratio between the population at (x, y) and that at (0, 0). In SI, we illustrate the derivation of this principle from the Fourier's law of heat conduction.

As for the population diffusion, k times t indicates the degree of diffusion within time t, which remains independent of particular cases. Therefore, C is primarily determined by the population size at the center (Q) and the distance between the center and the competing region (d). For the sake of simplicity and not losing generality, we set kt = 1, and assume that the competing languages were brought to the competing region only once. Now, equation **[2]** can be simplified as equation **[3]**:

$$C = \frac{Q}{(4\pi)^{\frac{3}{2}}} e^{-\frac{d^2}{4}} \quad \textbf{[3]}$$

Suppose that the impacts (σ_1 and σ_2) that competing languages (L_1 and L_2) cast upon each other are reflected by the population sizes of these languages and the distances between the competing region and the population centers of these languages, we have:

$$\begin{cases} \sigma_1 = \frac{Q_2}{Q_1} e^{\frac{d_1^2 - d_2^2}{4}} \\ \sigma_2 = \frac{Q_1}{Q_2} e^{\frac{d_2^2 - d_1^2}{4}} \end{cases} \quad \textbf{[4]}$$

Here, d_1 denotes the Euclidean distance from the competing region to the population center of L_1, and d_2 the Euclidean distance from the competing region to the center of L_2. If the competing region lies in the center of L_i, $d_i = 0$.

Language inheritance principle I. In our model, the inheritance rate of a language (r) reflects the inheritance capacity of this language during learning. In biology, during reproduction, the species with high inheritance capacity tend to proliferate in future generations, whereas the species with low inheritance capacity may gradually become extinct in future generations. Likewise, during language learning, a language with a high inheritance rate tends to be widely learned by language learners, whereas a language with a low inheritance rate may not be less preferred by language learners.

Noting these similarities between language learning and biological reproduction and between the inheritance capacity of language and that of species, we propose the language inheritance principle I, based on the genetic inheritance principle, to calculate language inheritance rates. To be specific, this principle is derived from the Hardy-Weinberg principle in genetics [24, 25], which states that without disturbing influences, both allele and genotype frequencies in a population remain constant across generations. some of the disturbing influences include: non-random mating, limited population size, mutation or migration of alleles in or between populations, selection for or against certain genotypes, genetic drift or flow, and others. Likewise, the language inheritance principle I states that populations speaking different languages also remain constant across generations in an ideal condition, where: (a) the global population is infinite or sufficiently large; (b) the new generation learns each language randomly, and masters one language at least and two at most (with refinement, this principle also works in tri- or multi-lingual situations); and (c) there is no sudden change of language, birth of new language, or selective pressure for or against any language. In SI, we give the proof of this principle.

Following this principle, we can approximate the occurring probabilities of competing languages in the new generation, which echo the inheritance rates of these languages (r_1 and r_2). For example, referring to the questionnaires about informants' language choice, we can obtain the basic information from informants, including their names, genders, ages, primary and secondary languages, based on which we can calculate the type frequencies of involved languages. Equation **[5]** shows the formulas in the case involving two languages (A and B):

$$\begin{aligned} p(AA) &= \frac{n_1}{n_1 + n_2 + n3} \\ p(AB) &= \frac{n_2}{n_1 + n_2 + n_3} \\ p(BB) &= \frac{n_3}{n_1 + n_2 + n_3} \end{aligned} \quad \textbf{[5]}$$

Here, n_1, n_2, and n_3 are the numbers of monolingual speakers of A, bilingual speakers, and monolingual speakers of B, respectively. Then, we can estimate the occurring frequencies of these languages (r_A and r_B), namely the

inheritance rates of the population speaking these languages (r_1 and r_2), as in equation **[6]**:

$$\begin{cases} r_1(r_A) = p(A) = p(AA) + 0.5p(AB) \\ r_2(r_B) = p(B) = p(BB) + 0.5p(AB) \end{cases} \quad [6]$$

Language inheritance principle II. In practice, the language survey data may not be sufficient or available at all. In fact, lacking sufficient data is a typical situation in empirical research. In this situation, the traditional way of using a large amount of data to fit parameter values becomes infeasible. In order to expand the application scope of our language competition model, we propose the language inheritance principle II to estimate the inheritance rates of competing languages in cases that lack sufficient direct data.

This principle is inspired by the well-attested lexical diffusion dynamics in computational linguistics [26, 27]. This dynamics, derived from the epidemic model [28], uses a logistic curve to describe lexical diffusion, as in equation **[7]**:

$$p(t) = \frac{\varepsilon e^{\alpha t}}{1 + \varepsilon(e^{\alpha t} - 1)} \quad [7]$$

Here, $p(t)$ calculates the proportion of the population using the changed lexical form, and $\varepsilon = p(t_0)$. When two individuals respectively using the changed and unchanged lexical forms contact, α denotes the probability for the individual using the unchanged form to start using the changed form.

The logistic curve was originally proposed to describe population growth [29], and α denoted the proportional increase in the population within a unit of time. In lexical diffusion, this curve was adopted to describe the changes in populations using different types of lexical forms, and α helped adjust the speed of lexical diffusion. As for language competition, the inheritance rates of populations speaking competing languages resembles the proportions of populations using changed lexical forms in lexical diffusion. Therefore, these inheritance rates can also be described by logistic curves, in which α helps adjust the speed of competition:

$$r(t) = \frac{\varepsilon e^{\alpha t}}{1 + \varepsilon(e^{\alpha t} - 1)} \quad [8]$$

We assume that both the population sizes of competing languages and the geographical distances between the population centers of these languages and the competing region collectively affect competition. In order to let α reflect these factors, we adopt equation **[8]** to calculate α, as in equation **[9]**:

$$\alpha = C = \frac{Q}{(4\pi)^{\frac{3}{2}}} e^{-\frac{d^2}{4}} \quad [9]$$

If the competing languages were brought to the competing region at the initial state ($t = 0$). After a unit of time, at $t = 1$, the influences of the population centers of those competing languages on language learning in the competing region start to take effect, and the inheritance rates of these languages can be estimated as in equation **[10]**, where ε is set according to particular cases:

$$\begin{cases} r_1 = r_A(1) = \frac{\varepsilon e^{\alpha_A}}{1+\varepsilon(e^{\alpha_A}-1)} \\ r_2 = r_B(1) = \frac{\varepsilon e^{\alpha_B}}{1+\varepsilon(e^{\alpha_B}-1)} \end{cases} \quad [10]$$

Evaluating procedure and evaluating indices. In a real case of language competition within a particular time period, we adopt the following procedure to evaluate our language competition model. First, we set the monolingual population data at the starting time step as the initial state of the model, and then, calculate the language impacts and inheritance rates following the above-mentioned principles. After obtaining the estimated parameter values, we let our language competition model predict the monolingual population sizes at the later time steps, and then, compare the predicted data with the empirical population data in that case.

In order to compare the predicted data with the empirical data, we define mean square error (MSE) and normalized mean square error ($NMSE$) as in equation **[11]**:

$$MSE = \sqrt{\frac{\sum_i (x_{pred(i)} - x_{real(i)})^2}{n}}, NMSE = \frac{MSE}{N} \quad [11]$$

Here, $x_{pred(i)}$ is the ith predicted data of the competition model, $x_{real(i)}$ is the corresponding empirical data, $N = N_1 = N_2$, and n is the number of time points where the empirical data are available. Note that the data in the initial state are excluded, since the error of that state is 0.0.

Results

We use four cases of language competition to evaluate our language competition model and relevant principles. The first two cases, namely the English-Welsh competition in Wales, UK and the English-Gaelic competition in Scotland, UK, contain sufficient empirical data. These cases illustrate the reliability of our model in replicating the historical data of language competition. In order to exclude the possible dependence on the initial time step, we also take the empirical data at different time steps as the initial states and further prove that the predicted data in these situations also largely match the empirical data. The last two cases, namely the English-Mandarin competition and the Mandarin-Malay competition in Singapore, do not contain many direct data, especially the exact numbers of monolinguals and bilinguals. In these cases, we have to use the language inheritance principle II to estimate the inheritance rates of competing languages, but the language competition model with thus-estimated parameter values still reliably replicate the limited amount of the empirical data. These cases illustrate the applicability of our model especially in cases lacking sufficient empirical data. In the following calculations, the shown values are rounded to 3 decimal places.

The English-Welsh competition in Wales, UK. This competition took place around the 20th century in Wales, UK. Within a century, the number of monolingual speakers of Welsh diminished severely, and many local people became English-Welsh bilinguals or English monolinguals [30]. The historical data tracing this competition from 1901 to 2001 were available (see Table S2 in SI).

Following the evaluation procedure, we set $x_1(0) = Q_1 = 1.029$, $x_2(0) = Q_2 = 0.309$ (in millions), according to the data in 1901. We set $N_1 = N_2 = 2.299$ (in millions), which was the sum of the English and Welsh monolingual populations in 2001. Since the competition occurred primarily in Wales, we set $Q_1 = 1.029$, $Q_2 = 0.309$ (in millions) according to the population data in 1901, and $d_1 = d_2 = 0$. Then, following the language diffusion principle (equation **[2]**), we calculate the impacts of English and Welsh (σ_1 and σ_2):

$$\begin{cases} \sigma_1 = \frac{Q_2}{Q_1} e^{\frac{d_1^2 - d_2^2}{4}} = \frac{0.309}{1.029} = 0.300 \\ \sigma_2 = \frac{Q_1}{Q_2} e^{\frac{d_2^2 - d_1^2}{4}} = \frac{1.029}{0.309} = 3.330 \end{cases} \quad [12]$$

Meanwhile, based on the data in 1901 and following the language inheritance principle I (equation **[6]**), we calculate the inheritance rates of populations speaking English and Welsh (r_1 and r_2):

$$\begin{cases} r_1 = p(AA) + 0.5p(AB) = 0.501 + 0.5 \times 0.348 = 0.675 \\ r_2 = p(BB) + 0.5p(AB) = 0.151 + 0.5 \times 0.348 = 0.325 \end{cases} \quad [13]$$

Now, based on MSE and $NMSE$ (equation **[11]**, where $n = 18$, covering all the data points except in 1901 in Table S2), we obtain the best solution of the differential equations in our competition

model (see SI text and Figure S1(a)). Figure 1(a) shows the predicted data of this solution and the corresponding historical data. This figure and MSE (0.068 (in millions)) or $NMSE$ (2.945%) collectively indicate that by estimating its parameters following the proposed principles, our language competition model reliably replicate the historical data of this competition.

The English-Gaelic competition in Scotland, UK. This competition took place in the Sutherland area of Scotland, UK, also around the 20th century, and resulted in a quick disappearance of Gaelic monolinguals [31]. The historical data tracing this competition from 1891 to 1971 were available (see Table S3 in SI). Like the English-Welsh competition, we set $x_1(0) = Q_1 = 5.804$, $x_2(0) = Q_2 = 1.094$ (in thousands) according to the data in 1891, $N_1 = N_2 = 11.185$ (in thousands) according to the sum of the English and Gaelic monolingual populations in 1971, and $d_1 = d_2 = 0$. Then, the impacts of English and Gaelic are, $\sigma_1 = 5.305$ and $\sigma_2 = 0.188$, and the inheritance rates are, $r_1 = 0.612$ and $r_2 = 0.388$. We obtain the best solution (see SI text and Figure S1(b)) based on MSE and $NMSE$ (equation **[11]**, where $n = 14$, covering all the 14 data points except in 1891 in Table S3), and illustrate the predicted data of this solution and the historical data in Figure 1(b). This figure and MSE (0.352 (in thousands)) or $NMSE$ (3.147%) also reveal a good match between the predicted data and the historical data in this case.

In these two cases, also discussed elsewhere [8], if a year other than 1901 or 1891 is set as the initial state of the model, we need to re-calculate the parameters according to the historical data in that year, let the model with these new parameter values predict the population data in the remaining time steps, and compare the predicated data with the corresponding historical data. Figures S2 and S3 in SI show that the predicted data under different initial states still match the historical data very well. These results indicate that the reliable replication of the empirical data based on our model is not dependent on particular time steps.

The English-Mandarin competition in Singapore. In Singapore, the majority of population are immigrants. As a former colony of UK, the English in Singapore is under great influence from UK, whereas Mandarin was brought to Singapore primarily by immigrants from Fujian and Guangdong in China, and Malaysian from Malaysia form the Malay speaking population. English, Mandarin and Malay are now all official languages there, and competitions among them are very frequent, especially at home. Noting these, we focus our study on the predominant household language (the language or dialect spoken by the majority of household members) and the most frequently spoken language at home (the language or dialect that a person uses frequently at home when speaking to household) [32]. In these cases, we lack direct data as those in the above two cases. Nonetheless, based on the language diffusion principle and the language inheritance principle II, we can still calculate the parameter values and make reasonable predictions about these competitions.

As for the English-Mandarin competition to be the predominant household language, we set the competition time period from 1985 to 2010, based on the limited availability of the empirical data. We set London as the population center of English, and the geographical center of Fujian and Guangdong as the population center of Mandarin, then, $d_1 = 1.886$ (the distance from London to Singapore), $d_2 = 0.800$ (the distance from the Mandarin center to Singapore) (in 10^4 kilometers) (here, using 10^4 kilometers as the distance scale is to confine the calculated impact values within the same magnitude as those in the above cases). In 1985, the population of UK was 56.550 million (according to the population census of UK, http://www.tradingeconomics.com/united-kingdom/population), the populations of Fujian and Guangdong in China were 27.130 million and 62.530 million, respectively (according to the population census of China in 1985 [33]), so the total population was 89.660 million. Accordingly, we set $Q_1 = 56.550$ and $Q_2 = 89.660$ (in millions). Then, following the language diffusion principle, we can calculate the impacts of English and Mandarin (σ_1 and σ_2):

$$\begin{cases} \sigma_1 = \frac{Q_2}{Q_1} e^{\frac{d_1^2 - d_2^2}{4}} = \frac{89.660}{56.550} e^{\frac{1.886^2 - 0.800^2}{4}} = 3.286 \\ \sigma_2 = \frac{Q_1}{Q_2} e^{\frac{d_2^2 - d_1^2}{4}} = \frac{56.550}{89.660} e^{\frac{0.800^2 - 1.886^2}{4}} = 0.304 \end{cases} \quad \textbf{[14]}$$

We adopt the language inheritance principle II (equation **[10]**) to estimate the inheritance rates. In these two cases in Singapore, we set $\varepsilon = 0.1$. As for α, we calculate it following equation **[9]**:

$$\begin{cases} \alpha_A = C_A = \frac{Q_A}{(4\pi)^{\frac{3}{2}}} e^{-\frac{{d_A}^2}{4}} = \frac{56.550}{(4\pi)^{\frac{3}{2}}} e^{-\frac{1.886^2}{4}} = 0.522 \\ \alpha_B = C_B = \frac{Q_B}{(4\pi)^{\frac{3}{2}}} e^{-\frac{{d_B}^2}{4}} = \frac{89.660}{(4\pi)^{\frac{3}{2}}} e^{-\frac{0.800^2}{4}} = 1.715 \end{cases} \quad \textbf{[15]}$$

The inheritance rates of English and Mandarin (r_1 and r_2) are:

$$\begin{cases} r_1 = r_A(1) = \frac{0.1 e^{0.522}}{1 + 0.1(e^{0.522} - 1)} = 0.158 \\ r_2 = r_B(1) = \frac{0.1 e^{1.715}}{1 + 0.1(e^{1.715} - 1)} = 0.382 \end{cases} \quad \textbf{[16]}$$

As for this competition, we only have the population data at 1985, 2000, 2005, and 2010 (see Table S4 in SI). We set $x_1(0) = 0.227$, $x_2(0) = 0.201$ (in millions) according to the data in 1985 as the initial state of the model. We set $N_1 = N_2 = 5.077$ (in millions), according to the total population of Singapore in 2010. Then, we obtain the best solution (see SI text and Figure S1(c)) based on MSE and $NMSE$ (equation **[11]**, where $n = 6$, covering the six data points except in 1985 in Table S4), and show the predicted data of this solution and the historical data in Figure 2(a). This figure and MSE (0.028 (in millions)) or $NMSE$ (0.558%) reveal a good match between the predicted data and the limited amount of empirical data within this time period.

The Mandarin-Malay competition in Singapore. As regards the Mandarin-Malay competition to be the most frequently spoken language at home, we set the competition period from 1980 to 2010 based on our available data (see Table S5 in SI). We set the geographical center of Fujian and Guangdong as the population center of Mandarin, and Kuala Lumpur, the capital of Malaysia, as the population center of Malaysian people, then, $d_1 = 0.800$ (the distance from the Mandarin center to Singapore), $d_2 = 0.030$ (the distance from Kuala Lumpur to Singapore) (in 10^4 kilometers). In 1980, the Chinese population was 77.460 millions (the sum of the populations in Fujian and Guangdong in the population census of China in 1980 [34]), the Malaysian population in Malaysia was 13.763 millions (http://www.tradingeconomics.com/malaysia/population). Accordingly, we set $Q_1 = 77.460$ and $Q_2 = 13.763$ (in millions). Then, following the language diffusion principle, we calculate the impacts of Mandarin and Malay: $\sigma_1 = 0.209$, $\sigma_2 = 4.797$. Following the inheritance principle II, we calculate the inheritance rates of Mandarin and Malay: $r_A = 0.328$, $r_B = 0.131$.

For this competition, we only have the population data at 1980, 1990, 2000, and 2010 (see Table S5 in SI). We set $x_1(0) = 0.233$, $x_2(0) = 0.317$ (in millions) according to the data in 1980 as the initial state of the model. We set $N_1 = N_2 = 5.077$ (in millions), according to the total population of Singapore in 2010. Then, we obtain the best solution (see SI text and Figure S1(d)) based on MSE and $NMSE$ (equation **[11]**, where $n = 6$, covering the six data points except in 1980 in Table S5), and show the predicted data of this solution and the historical data in Figure 2(b). This figure and MSE (0.159 (in millions)) or $NMSE$ (3.125%) also reveal a good match between the predicted data and the limited amount of empirical data within this time period.

Discussion and conclusion

Reasonably defining and accurately estimating key parameters are important criteria for evaluating mathematical models of real world phenomena, yet both aspects have not been explicitly addressed in many models of language competition, i.e. the sole parameter, prestige, cannot clearly address the influence of many factors that affect language competition. A more realistic model should define concrete parameters that denote these factors and compute their values based on the relevant data that reflect the influences of these factors. To this purpose, we define language impacts and inheritance rates as the key parameters for language competition, and propose three principles that link these parameters with population sizes of competing languages, geographical distances between populations, and speaker distributions in competing regions, which allows explicitly calculating the values of these parameters from data of population censuses and language surveys. This approach greatly extends not only the reusability of available linguistic resources obtained from linguistic field works, but also the applicability of the language competition model incorporating these meaningful parameters, especially in cases lacking sufficient competition data.

Our study also bears important guidance for future modeling exploration of language competition. On the one hand, in the language diffusion principle, we adopt the equation of heat diffusion to describe the diffusion of populations of competing languages, and the good match between the predicted data and the historical data reveals the intrinsic commonness between these social and physical phenomena. In the language inheritance principle I, we never neglect bilinguals when calculating the impacts of competing languages; otherwise, the model would never replicate the English-Welsh and English-Gaelic competitions, since bilinguals in both cases used to take up sufficiently big proportions in total populations. In the language inheritance principle II, we modify the lexical diffusion dynamics and apply it to estimate the inheritance rates of competing languages in cases lacking sufficient data of language surveys.

On the other hand, we derive the language competition model from the classic ecological system dynamics in evolutionary biology. This dynamics resembles language competition in many aspects, and factors affecting it also have their linguistic correspondences and may cast similar effects on language competition. In addition, this model highlights the roles of population size and geographical distance in language competition, which have been noticed very recently in some empirical and simulation studies [10, 12, 13, 35, 36, 37]. Furthermore, the language diffusion principle can be directly applied in a 3D world, which allows more systematic connection with the geographical information systems, more realistic simulation of geographical barriers, and more accurate prediction of language competition in various geographical conditions.

All these aspects collectively indicate that systematically linking similar linguistic phenomena (e.g. lexical diffusion and general language competition) and adopting (with necessary modifications) well-attested theories, models, methods, and data from physics, geography, population genetics, and evolutionary biology into linguistics research are efficient ways to obtain more insightful understanding of linguistic phenomena, as already practiced in many recent studies [38, 39, 40].

ACKNOWLEDGMENTS. This work is funded by the National Social Science Research Foundation of China (No. 09JZD0007 and 12CYY031), the Ministry of Education of China (No. 12YJC740082), and the Seed Funding Programme for Basic Research of the University of Hong Kong.

1. M.A. Nowak, N.L. Komarova, and P. Niyogi, *Computational and evolutionary aspects of language*, Nature, 417 (2002), pp. 611–617.
2. E. Lieberman, J-B. Michel, J. Jackson, T. Tang, and M.A. Nowak, *Quantifying the evolutionary dynamics of language*, Nature, 449 (2007), pp. 713–716.
3. S. Wichmann, *The emerging field of language dynamics*, Lang. Ling. Comp., 2 (2008), pp. 442–455.
4. R.V. Solé, B. Corominas-Murta, and J. Fortuny, *Diversity, competition, extinction: The ecophysics of language change*, J. R. Soc. Interface, 7 (2010), pp. 1647–1664.
5. N. Kandler, *Demography and language competition*, Hum. Biol., 81 (2009), pp. 181–210.
6. J. Mira, L.F. Seoane, and J.J. Nieto, *The importance of interlinguistic similarity and stable bilingualism when two languages compete*, New J. Phys., 13 (2011), 033007.
7. I. Baggs and H.I. Freedman, *A mathematical model for the dynamics of interactions between a unilingual and a bilingual population: Persistence versus extinction*, J. Math. Sociol., 16 (1990), pp. 51–75.
8. D.M. Abrams and S.H. Strogatz, *Modeling the dynamics of language death*, Nature, 424 (2003), pp. 900.
9. J. Mira and A. Paredes, *Interlinguistic similarity and language death dynamics*, Europhys. Lett., 69 (2005), pp. 1031–1034.
10. D. Stauffer and C. Schulze, *Microscopic and macroscopic simulation of competition between languages*, Phys. Life Rev., 2 (2005), pp. 89–116.
11. J.M. Minett and W.S-Y. Wang, *Modeling endangered languages: The effects of bilingualism and social structure*, Lingua, 118 (2008), pp. 19–45.
12. V.M. De Oliveira, M.A. Gomes, and I.R. Tsang, *Theoretical model for the evolution of linguistic diversity*, Physica A., 361 (2006), pp. 361–370.
13. M. Patriarca and E. Heinsalu, *Influence of geography on language competition*, Physica A, 388 (2009), pp. 174–186.
14. F. Vazquez, X. Castelló, and M. San Miguel, *Agent based models of language competition: Macroscopic descriptions and order-disorder transitions*, J. Stat. Mech-Theory Exp., 4 (2010), P04007.
15. X. Castelló, V.M. Eguíluz, and M. San Miguel, *Ordering dynamics with two non-excluding options: Bilingualism in language competition*, New J. Phys., 8 (2008), pp. 308–324.
16. S.G. Thomason and T. Kaufman, *Language contact, creolization, and genetic linguistics*, 1988.
17. D. Crystal, *Language death*, 2000.
18. S.S. Mufwene, *The ecology of language evolution*, 2001.
19. W. Labov, *Principles of linguistic change; Social factors*, 2001.
20. S.C. Manrubia, J.B. Axelsen, and D.H. Zanette, *Role of demographic dynamics and conflict in the population-area relationship for human languages*, PLoS ONE, 7 (2012), e40137.
21. A.J. Lotka, *Contribution to the theory of periodic reaction*, J. Phys. Chem., 14 (1910), pp. 271–274.
22. V. Volterra, *Lecons sur la theorie mathematique de la lutte pour la vie*, 1931.
23. M.E. Gilpin and F.J. Ayala, *Global models of growth and competition*, Proc. Nat. Acad. Sci. USA, 70 (1973), pp. 3590–3593.
24. G.H. Hardy, *Mendelian proportions in a mixed population*, Science, 28 (1908), pp. 49–50.
25. W. Weinberg, *Über den Nachweis der Vererbung beim Menschen*, Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg, 64 (1908), pp. 368–382.
26. Z-W. Shen, *Exploring the dynamic aspect of sound change*, 1997.
27. W.S-Y. Wang and J.W. Minett, *The invasion of language: Emergence, change and death*, Trends Ecol. Evol., 20 (2005), pp. 263–269.
28. F. Brauer and C. Castillo-Chavez ,*Mathematical models in population biology and epidemiology*, 2001.
29. S.E. Kingsland, *Modeling nature: Episodes in the history of population ecology*, 1995.
30. J.W. Aitchison and H. Carter, *Language, economy, and society: The changing fortunes of the Welsh language in the twentieth century* , 2000.
31. C.W.J. Withers, *Gaelic in Scotland 1691-1981: The geographical history of a language* , 1984.
32. B.G. Leow, *Census of population 2000: Education, language and religion*, 2000.
33. National Bureau of Statistics of China, *China statistical yearbook 1985*, 1986.
34. National Bureau of Statistics of China, *China statistical yearbook 1980*, 1981.
35. M. Patriarca and T. Leppanen, *Modeling language competition*, Physica A, 338 (2004), pp. 296–299.
36. E.W. Holman, C. Schulze, D. Stauffer, and S. Wichmann, *On the relation between structural diversity and geographical distance among languages: Observations and computer simulations*, Ling. Typ., 11 (2007), pp. 393–421.
37. C. Schulze and D. Stauffer, *Competition of languages in the presence of a barrier*, Physica A, 379 (2007), pp. 661.
38. T.F. Jaeger, P. Graff, W. Croft, and D. Pontillo, *Mixed effect models for genetic and areal dependencies in linguistic typology*, Ling. Typ., 15 (2011), pp. 281–319.
39. T. Gong, L. Shuai, M. Tamariz, and G. Jaeger, *Studying language change using Price equation and Pólya-urn dynamics*, PLoS ONE, 7 (2012), e33171.
40. S.C. Levinson and R.D. Gray, *Tools from evolutionary biology shed new light on the diversification of languages*, Trends Cogn. Sci., 16 (2012), pp. 167–173.

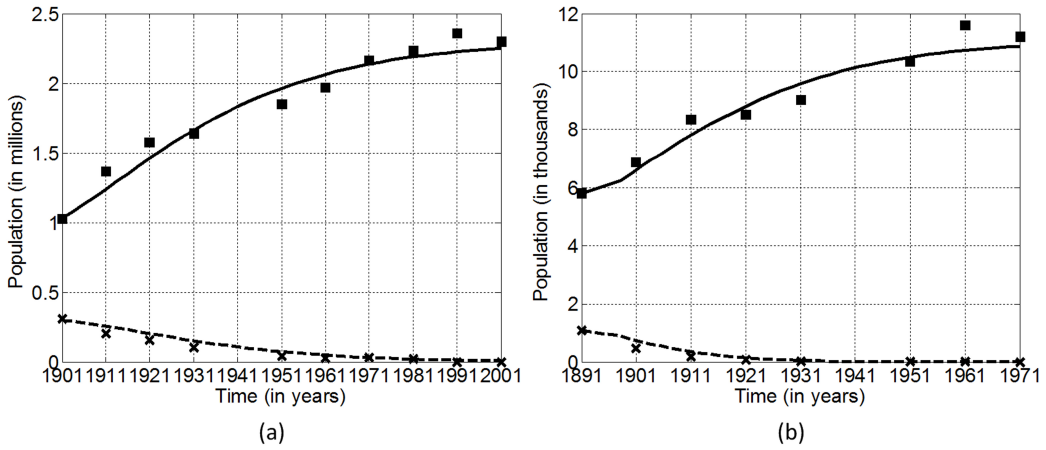

Fig. 1. The predicted data of the best solutions and the historical data in the English-Welsh competition from 1901 to 2001 (a) and the English-Gaelic competition from 1891 to 1971 (b). Solid lines: predicted data of English monolingual populations. Dash lines: predicted data of Welsh or Gaelic monolingual populations. Squares: historical data of English monolingual populations. Crosses: historical data of Welsh monolingual populations.

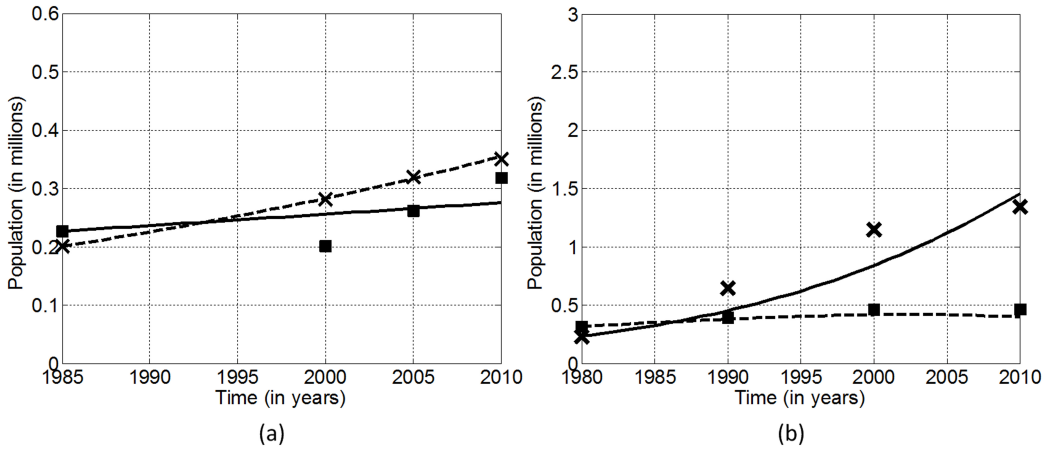

Fig. 2. The predicted data of the best solutions and the historical data in the English-Mandarin competition from 1985 to 2010 (a) and the Mandarin-Malay competition from 1980 to 2010 (b). Solid lines: predicted data of English (a) or Mandarin (b) monolingual populations. Dash lines: predicted data of Mandarin (a) or Malay (b) monolingual populations. Squares in (a): historical data of English monolingual populations in years 1985, 2000, 2005, 2010. Squares in (b): historical data of Mandarin monolingual populations in years 1980, 1990, 2000, 2010. Crosses in (a): historical data of Mandarin monolingual populations in years 1985, 2000, 2005, 2010. Crosses in (b): historical data of Malay monolingual populations in years 1980, 1990, 2000, 2010.